Multinomial logit random effects models

Jonathan Hartzel, Alan Agresti and Brian Caffo

Statistical Modelling 2001 1: 81
DOI: 10.1177/1471082X0100100201

Statistical Modelling 2001; **1**: 81–102

Multinomial logit random effects models

Jonathan Hartzel[1], Alan Agresti[2] and Brian Caffo[2]
[1]Merck Research Labs, West Point, USA
[2]Department of Statistics, University of Florida, Gainesville, USA

Abstract: This article presents a general approach for logit random effects modelling of clustered ordinal and nominal responses. We review multinomial logit random effects models in a unified form as multivariate generalized linear mixed models. Maximum likelihood estimation utilizes adaptive Gauss–Hermite quadrature within a quasi-Newton maximization algorithm. For cases in which this is computationally infeasible, we generalize a Monte Carlo EM algorithm. We also generalize a pseudo-likelihood approach that is simpler but provides poorer approximations for the likelihood. Besides the usual normality structure for random effects, we also present a semi-parametric approach treating the random effects in a non-parametric manner. An example comparing reviews of movie critics uses adjacent-categories logit models and a related baseline-category logit model.

Key words: adjacent-categories logit; baseline-category logit; generalized linear mixed model; nominal variable; non-parametric maximum likelihood; ordinal variable; quasi symmetry

Data and software available from: http://stat.uibk.ac.at/SMIJ

Received February 2001; **revised** June 2001; **accepted** July 2001

1 Introduction

In many studies data occur in clusters, such as a cluster of repeated measurements for each subject in a study. One approach to modelling such data includes random effects for the subjects or clusters in the linear predictor. This provides a mechanism of accounting for certain correlation structures among the clustered observations and for overdispersion among discrete responses modelled with ordinary Poisson or binomial models.

When the response has distribution in the exponential family, generalized linear mixed models (GLMMs) extend generalized linear models (GLMs) to contain random effects. In GLMMs, the response distribution is defined conditionally on the random effects. The random effects are commonly assumed to be multivariate normal, although this is not necessary. For instance, an alternative approach uses a conjugate distribution (Lee and Nelder, 1996). With normal random effects the marginal distribution of the response, obtained by integrating out the random effects, does not have closed form. Numerical integration using Gauss–Hermite quadrature (e.g., Hinde, 1982; Anderson and Aitkin, 1985) or Monte Carlo techniques (e.g., Zeger and Karim, 1991; McCulloch, 1994; McCulloch, 1997; Booth and Hobert, 1999) or approximation methods such as the Laplace approximation and Taylor series expansions (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) are used to

Address for correspondence: A Agresti, Department of Statistics, University of Florida, Gainesville, FL 32611-8545, USA. E-mail: aa@stat.ufl.edu

0962-2802(01)ST009RA

approximate the marginal distribution and hence the likelihood function and ultimately the maximum likelihood (ML) estimates and their standard errors.

Numerous articles have developed GLMMs for (conditionally) binomial (e.g., Stiratelli *et al.*, 1984; Anderson and Aitkin, 1985; Gilmour *et al.*, 1985) and Poisson (e.g., Hinde, 1982; Breslow, 1984) distributed responses. There has been less development for multinomial responses, with the majority of the research focused on ordinal models with logit and probit link functions for cumulative probabilities. Harville and Mee (1984) proposed a cumulative probit random effects model that utilized Taylor series approximations for intractable integrals. Jansen (1990) and Ezzet and Whitehead (1991) proposed random intercept cumulative probit and logit models, respectively, and employed quadrature techniques. More general ordinal random effects models that allowed multiple random effects were proposed by Hedeker and Gibbons (1994), who applied Gauss–Hermite quadrature within a Fisher scoring algorithm, and Tutz and Hennevogl (1996), who used quadrature and Monte Carlo EM algorithms. Other links have received less attention. Ten Have and Uttal (1994) used a continuation-ratio logit random effects model for analysing multiple discrete time survival profiles. See also Coull and Agresti (2000) and Dos Santos and Berridge (2000). Crouchley (1995) developed models with cumulative complementary log-log link using random effects distributions for which the marginal distribution has closed form. He and Ten Have (1996) extended a binary random effects model of Conaway (1990) to ordinal data; a closed-form expression for the likelihood function results from using the cumulative complementary log-log link and a log-gamma random effects distribution.

The modelling of nominal responses with random effects has received less attention, and most of that is in the econometric or psychometric literature. A potential application is when subjects respond to several related multiple choice questions in which the response categories are unordered or not fully ordered. Item response versions of such a model generalize the Rasch model (Rasch, 1961; Adams and Wilson, 1996; Adams *et al.*, 1997; Bock, 1972). Bock (1972) described models more general than most item response models in allowing the random effects variance to vary across items. An appropriate link function for nominal responses is the baseline-category logit. Daniels and Gatsonis (1997) used this in a hierarchical Bayesian model for applications in health services research. Revelt and Train (1998) introduced random effects in discrete choice models, with random coefficients for the explanatory variables which, in such models, may themselves vary according to the response category. For related applications, see Chintagunta *et al.* (1991), Jain *et al.* (1994), and Gönül and Srinivasan (1993).

This article discusses a general approach for logit random effects modelling of clustered multinomial (ordinal or nominal) responses. Following Tutz and Hennevogl (1996), we present models as multivariate generalized linear mixed models. Our main purpose is to survey and unify existing logit random effects models for multicategory responses. In addition, we extend models and estimation methods. For instance, for ordinal responses we consider adjacent-categories logit random effects models. These do not seem to have been considered previously, except in a companion paper (Hartzel *et al.*, 2001) for a particular application. Also, for nominal responses we extend the baseline-category logit model by allowing general correlation structure for the random effects.

Our estimation approach extends those considered previously for multinomial data, both in approximating the likelihood function and in the treatment of the random effects distribution. To obtain ML estimates, we utilize adaptive Gauss–Hermite quadrature

(Liu and Pierce, 1994) within a quasi-Newton maximization algorithm. However, this approach is computationally infeasible when the integral dimension is large. For such cases we generalize a Monte Carlo EM algorithm proposed by Booth and Hobert (1999). We also present a generalization of pseudo-likelihood(PL) approaches (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) that are simpler but provide poorer approximations.

Besides the usual normality structure for random effects, we present an approach that treats random effects in a non-parametric manner. This deals with concerns of bias due to misspecification of that distribution. We also mention a limited simulation study designed to investigate effects of such misspecification.

Section 2 defines the general multinomial random effects model and mentions special cases for nominal and ordinal responses. Section 3 deals with model-fitting, presenting three ways of approximating the likelihood function, and discusses model inference and prediction. Section 4 discusses the non-parametric random effects approach. Section 5 illustrates the adjacent-categories and baseline-category logit link for an example in which several movie critics reviewed a sample of movies. That section also discusses a connection between multinomial logit random effects models for within-subjects effects in contingency tables and corresponding loglinear models.

2 Model specification

Suppose cluster i has T_i categorical observations. Let Y_{ij} denote the jth observation in cluster $i, j = 1, \ldots, T_i$, with response probabilities $\{\pi_{ijr} = P(Y_{ij} = r), r = 1, \ldots, R\}$. Let $\mathbf{x}_{ij}$ denote a column vector of explanatory variables for that observation, and let $\mathbf{w}_{ij}$ denote a column vector of coefficients for the random effects.

2.1 Logits for nominal responses

When the categories for Y_{ij} are unordered, logits pair each response category with an arbitrary baseline category, say category R. Including cluster-specific random effects $\mathbf{u}_{ir}$ for each logit

$$\log\left(\frac{\pi_{ijr}}{\pi_{ijR}}\right) = \alpha_r + \mathbf{x}'_{ij}\boldsymbol{\beta}_r + \mathbf{w}'_{ij}\mathbf{u}_{ir} \quad r = 1, \ldots, R-1 \tag{1}$$

is the *baseline-category logit* random effects model. The parameters α_r and $\boldsymbol{\beta}_r$ (and the random effects) depend on r, since the baseline category is arbitrary; with nominal responses there is no reason to expect effects to be similar for different r. The model definition is completed by specifying a distribution for $\mathbf{u}'_i = (\mathbf{u}'_{i1}, \ldots, \mathbf{u}'_{i,R-1})$. The usual random effects modelling approach treats $\{\mathbf{u}_i\}$ as independent multivariate normal variates. We recommend using an arbitrary covariance matrix $\boldsymbol{\Sigma}$. In particular, it is sensible to allow different variances for random effects that apply to different logits. With a common variance, that variance would not be the same as that for the implied random effect for a logit for an arbitrary pair of categories, logit$(\pi_{ijr}/\pi_{ijr'})$. With an arbitrary covariance structure the model is structurally the same regardless of the choice of baseline category.

The website www.uic.edu/~hedeker/mix.html has a FORTRAN program (MIXNO) by Hedeker for ML fitting of baseline-category logit models with random effects. It uses

Gauss–Hermite quadrature. See Hedeker (1999) for further details. In contrast to the general structure in (1) however, this assumes that $\{\mathbf{u}_{ir}\}$ are independent for different r or perfectly correlated. Since the random effects from different logits arise from the same subject, the assumption that they are independent may be unrealistic. However, to assume that they are perfectly correlated is too restrictive. Other software that can fit such models includes HLM (Bryk *et al.*, 2000), MLwiN (Goldstein *et al.*, 1998), and LIMDEP (Greene, 1998). LIMDEP also uses quadrature, whereas the other programs use approximations (such as Laplace approximations and Taylor series expansions) for the integral that determines the likelihood function. Unlike quadrature, these methods do not converge to ML when applied more finely. However, it is feasible to fit more complex models using them. Chen and Kuo (2001) discussed other software and connections with stratified proportional hazards models.

2.2 Logits for ordinal responses

A variety of ordinal regression models use the logit link. The most popular one is the *cumulative logit* model. A random effects version has form

$$\log\left(\frac{\pi_{ij1} + \cdots \pi_{ijr}}{\pi_{ij,r+1} + \cdots + \pi_{ijR}}\right) = \alpha_r + \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{w}'_{ij}\mathbf{u}_i \quad r = 1, \ldots, R-1 \tag{2}$$

In contrast to model (1), here the fixed effect $\boldsymbol{\beta}$ is the same for all logits, called the *proportional odds* assumption. More general models permit effects $\boldsymbol{\beta}_r$ to vary for different logits for at least some predictors, but have the structural problem whereby cumulative probabilities may be misordered. See, for instance, Hedeker and Mermelstein (1998). In model (2), each logit for the ith cluster also has the same random effect $\mathbf{u}_i$. Such simplifications result naturally from underlying latent variable models with the logistic distribution. The intercept parameters satisfy $\alpha_1 < \cdots < \alpha_{R-1}$. Tutz and Hennevogl (1996) considered a more general model that allowed differential random effects $\mathbf{u}_{ir}$ by logit, as in the baseline-category logit model. Estimation in this extended model is more complicated, since the intercepts must be reparameterized to ensure their ordering is not violated.

An alternative ordinal model uses logits, $\log(\pi_{ijr}/\pi_{ij,r+1})$, $r = 1, \ldots, R-1$. When this adjacent-categories logit (ACL) model has the same predictor form as model (2), $\boldsymbol{\beta}$ has log odds interpretations for all pairs of adjacent categories. The ACL model is more useful than the cumulative logit model when one wants descriptions to contrast probabilities of response in pairs of categories rather than above versus below various points on the response scale. Since intercepts in the ACL model are unordered, an extended model permitting different random effects for each logit does not require reparameterization and has the form

$$\log\left(\frac{\pi_{ijr}}{\pi_{ij,r+1}}\right) = \alpha_r + \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{w}'_{ij}\mathbf{u}_{ir} \quad r = 1, \ldots, R-1 \tag{3}$$

These logits are a basic set equivalent to the baseline-category logits. When $\boldsymbol{\beta}$ in (3) is replaced by $\boldsymbol{\beta}_r$, the ACL model is equivalent to a baseline-category logit model with an adjusted model matrix. The random effect $\mathbf{u}_{ir}$ in the baseline-category logit model then corresponds to $\mathbf{u}_{ir} + \cdots + \mathbf{u}_{i,R-1}$ from the ACL model. Thus, a simple covariance structure for the random effects in the ACL model implies a certain pattern for the covariance matrix for the random effects in the baseline-category logit model.

When the main interest is in estimating the fixed effects rather than characteristics of the intercept terms, the loss in parsimony in using the more general structure for the intercepts is usually unimportant. An advantage is that the model may fit much better. A disadvantage is that having a vector of random effects rather than a single random effect makes model fitting more challenging.

When used with a common fixed effect $\boldsymbol{\beta}$ and a common random effect $\mathbf{u}_i$ for each logit, ACL and cumulative logit models both specify location effects and imply stochastically ordered response distributions at different settings of predictors. They then typically provide similar substantive conclusions.

2.3 Multivariate generalized linear mixed models

Multinomial logit random effects models are a special case of a multivariate generalized linear mixed model (Tutz and Hennevogl, 1996), denoted by MGLMM. This formulation is advantageous for several reasons. As with GLMMs, MGLMMs provide a common framework using a unified notation and estimation method for a broad class of models, including known formulas for score functions and information matrices. We begin by defining the MGLMM and then embed the multinomial logit random effects model within its framework.

Let $\mathbf{y}_{ij}$ be a multivariate response vector for the jth observation in cluster i, $j = 1, \ldots, T_i$, $i = 1, \ldots, n$. With random effects $\mathbf{u}_i$, the conditional distribution $f(\mathbf{y}_{ij} \mid \mathbf{u}_i)$ belongs to the multivariate exponential family with

$$\boldsymbol{\mu}_{ij} = E(\mathbf{y}_{ij} \mid \mathbf{u}_i) = \mathbf{h}(\boldsymbol{\eta}_{ij}) \qquad \boldsymbol{\eta}_{ij} = \mathbf{Z}_{ij}\boldsymbol{\beta} + \mathbf{W}_{ij}\mathbf{u}_i \tag{4}$$

where $\boldsymbol{\beta}$ are fixed effects and $\mathbf{Z}_{ij}$ and $\mathbf{W}_{ij}$ are model matrices for the fixed and random effects. The function $\mathbf{h}(\boldsymbol{\eta}_{ij})$ is a vector of inverse link functions. The $\{\mathbf{u}_i\}$ are assumed to be independent from density function $g(\mathbf{u})$, such as the multivariate normal.

A given response may refer to a particular subject, or it may summarize counts for several subjects having the same random effect. For example, in multi-centre clinical trials regarding each centre as a cluster, observation j in centre i might refer to the multinomial summary for the n_{ij} subjects using the jth treatment in that centre. Thus, let $Y_{ij}^{(s)}$, $s = 1, \ldots, n_{ij}$, represent n_{ij} categorical outcomes for the jth observation of the ith cluster (where n_{ij} could equal 1). To express the multinomial model as a MGLMM, we first re-express each categorical response as a dummy response vector $\mathbf{y}_{ij}^{(s)\prime} = (y_{ij1}^{(s)}, \ldots, y_{ij,R-1}^{(s)})$ where

$$y_{ijr}^{(s)} = \begin{cases} 1 & \text{if } Y_{ij}^{(s)} = r \quad r = 1, \ldots, R-1 \\ 0 & \text{otherwise} \end{cases}$$

Thus, for n_{ij} independent repetitions, $\mathbf{y}_{ij} = \sum_{s=1}^{n_{ij}} \mathbf{y}_{ij}^{(s)}$ is distributed as multinomial with index n_{ij} and parameters $\boldsymbol{\pi}'_{ij} = (\pi_{ij1}, \ldots, \pi_{ij,R-1})$. Then $f(\bar{\mathbf{y}}_{ij} \mid \mathbf{u}_i)$, where $\bar{\mathbf{y}}_{ij} = \mathbf{y}_{ij}/n_{ij}$, is a member of the multivariate exponential family (Fahrmeir and Tutz, 2001, Chap. 3). The general multinomial model is defined by (4) in terms of the response vector $\bar{\mathbf{y}}_{ij}$. Specific cases result by specifying the inverse link function $\mathbf{h}(\boldsymbol{\eta}_{ij})$ and the model matrices.

For example, the baseline-category logit random effects model has

$$h_r(\boldsymbol{\eta}_{ij}) = \frac{\exp(\eta_{ijr})}{1 + \sum_{l=1}^{R-1} \exp(\eta_{ijl})} \qquad r = 1, \ldots, R-1$$

The canonical parameters equated to the linear predictor are $\log(\pi_{ijr}/\pi_{ijR}), r = 1, \ldots, R-1$. The fixed effects model matrix is then a $(R-1)$-row matrix with non-zero elements

$$\mathbf{Z}_{ij} = \begin{bmatrix} 1 & \mathbf{x}'_{ij} & & & & \\ & & 1 & \mathbf{x}'_{ij} & & \\ & & & & \ddots & \\ & & & & 1 & \mathbf{x}'_{ij} \end{bmatrix}$$

which are coefficients of $\boldsymbol{\beta}' = (\alpha_1, \boldsymbol{\beta}'_1, \ldots, \alpha_{R-1}, \boldsymbol{\beta}'_{R-1})$.

3 Model fitting and inference

We first assume multivariate normality with mean 0 and covariance matrix $\boldsymbol{\Sigma}$ for $\mathbf{u}$. Denote this density by $g(\mathbf{u}; \boldsymbol{\Sigma})$. The likelihood function for the multinomial model then has the form

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \mathbf{u}_i; \boldsymbol{\beta}) \right] g(\mathbf{u}_i; \boldsymbol{\Sigma}) \, \mathrm{d}\mathbf{u}_i \tag{5}$$

This section describes three approaches for approximating this integral. The first uses adaptive Gauss–Hermite quadrature. The second is an automated Monte Carlo EM algorithm that utilizes rejection sampling and estimates the Monte Carlo error to update the required Monte Carlo sample size. The third is a PL algorithm that utilizes simple Taylor series approximations.

3.1 Adaptive Gauss–Hermite, quasi-Newton algorithm

When feasible, in our opinion the best method for approximating the multivariate normal integrals is adaptive Gauss–Hermite quadrature (Liu and Pierce, 1994; Pinheiro and Bates, 1995). This method uses the same weights and nodes as Gauss–Hermite quadrature, but to increase efficiency it centres the nodes with respect to the mode of the function being integrated and scales them according to the estimated curvature at the mode. This dramatically reduces the number of quadrature points needed to approximate the integrals effectively. Although additional computing time is needed to compute the mode and curvature for each unique cluster, many fewer quadrature points are needed to obtain the same degree of accuracy. The number of unique clusters is the biggest factor in determining the extra amount of time adaptive quadrature requires. So, it is highly efficient for multinomial response data with categorical covariates, but may be slower for continuous covariates with large data sets, as the mode and curvature must be calculated for every cluster for every iteration. If computing time is a problem, we recommend using ordinary quadrature to get starting values with few quadrature points, and then use the adaptive version to improve accuracy.

For cluster i, we first calculate the mode $\boldsymbol{\omega}_i$ of the integrand $\prod_{j=1}^{T_i}[f(\bar{\mathbf{y}}_{ij} \mid \mathbf{u}_i; \boldsymbol{\beta})]g(\mathbf{u}_i; \boldsymbol{\Sigma})$ and centre the original Gauss–Hermite nodes about that point. Scaling the centred nodes according to the curvature of the integrand around the mode further localizes the quadrature points near the bulk of the integrand. An estimate of the curvature at the mode of the integrand results from inverting the negative of the second derivative matrix of the integrand evaluated at the estimated mode. We used numerical second derivatives for the estimation of the curvature, $\hat{\mathbf{Q}}_i$. Let m denote the dimension of $\mathbf{u}_i$ and denote the original Gauss–Hermite nodes by $\mathbf{z_l} = (z_{l_1}, \ldots, z_{l_m})$. Then, the adapted Gauss–Hermite nodes for the ith cluster are

$$\mathbf{z}^*_{i\mathbf{l}} = \hat{\boldsymbol{\omega}}_i + \sqrt{2}\, \hat{\mathbf{Q}}_i^{1/2}\mathbf{z_l}$$

where $\hat{\mathbf{Q}}_i^{1/2}$ results from the Cholesky decomposition of the estimated curvature $\hat{\mathbf{Q}}_i$.

Using the transformed Gauss–Hermite nodes for each cluster, the contribution to the likelihood by the ith cluster is approximated by

$$L_i(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \approx |\hat{\mathbf{Q}}|^{1/2}\, 2^{m/2} \sum_{\mathbf{l}} w_{\mathbf{l}} \left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \mathbf{z}^*_{i\mathbf{l}}; \boldsymbol{\alpha})\right] g(\mathbf{z}^*_{i\mathbf{l}}; \boldsymbol{\Sigma}) \exp(\mathbf{z}'_{\mathbf{l}}\mathbf{z}_{\mathbf{l}})$$

where $w_{\mathbf{l}} = \prod_{t=1}^{m} w_{l_t}$ and $\{w_{l_t}\}$ are the original Gauss–Hermite weights. Thus, the intractable integrals are approximated by a finite summation, with each of the m summations taken over K quadrature points. After approximating the likelihood, one can maximize it using a variety of methods. We developed a quasi-Newton algorithm using analytical first derivatives of the marginal log-likelihood and numerical second derivatives for estimating the Hessian matrix. At convergence, we computed and inverted the observed information matrix to obtain standard error estimates for the estimated model parameters. We recommend implementing this method by sequentially increasing the number of quadrature points K until the changes are negligible in both the estimates and standard errors.

3.2 Automated Monte Carlo EM algorithm

Adaptive quadrature with K points in each of m dimensions approximates the integrals with a finite summation over K^m quadrature points. Since the number of nodes increases exponentially in m, multivariate quadrature is currently computationally feasible only for integral dimensions up to about 5 or 6. For models with higher-dimensional integrals, more feasible methods for approximating integrals use Monte Carlo methods. In contrast to adaptive quadrature, Monte Carlo methods use K randomly sampled nodes $\mathbf{z}$ to approximate integrals. An important issue is then the number of nodes to sample to approximate adequately the integrals. Using too few nodes can result in a poor approximation to ML estimates due to Monte Carlo error, while too many nodes can significantly increase the computational time. Booth and Hobert (1999) proposed an automated Monte Carlo EM algorithm for estimation in GLMMs. At each iteration they assessed the Monte Carlo error in the current parameter estimates and increased the number of nodes if the error exceeded the change in the estimates from the previous iteration. To make this possible, independently and identically distributed random samples are generated at each iteration, allowing one to use standard Central Limit theory to assess the Monte Carlo error. We now outline an extension of their algorithm that, used with the EM algorithm, provides estimates and standard errors for multinomial logit random effects models.

The EM algorithm (Dempster *et al.*, 1977) is an iterative method for maximum likelihood estimation in situations with incomplete data. For MGLMMs, we treat the random effects $\mathbf{u}' = (\mathbf{u}_1', \ldots, \mathbf{u}_n')$ as the missing data and the complete-data log-likelihood as

$$l_C(\boldsymbol{\Psi}) = \log f\{\mathbf{y}, \mathbf{u}; \boldsymbol{\Psi}\} = \sum_{i=1}^{n}\left[\log\left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \mathbf{u}_i; \boldsymbol{\beta})\right] + \log g(\mathbf{u}_i; \boldsymbol{\Sigma})\right] \tag{6}$$

where $\boldsymbol{\Psi}' = (\boldsymbol{\beta}', \boldsymbol{\Sigma})$. In the E (expectation) step at iteration $(s+1)$, the expectation of the complete log-likelihood (6) is determined with respect to the conditional distribution $h(\mathbf{u} \mid \boldsymbol{\Psi}; \mathbf{y})$. That is,

$$\begin{aligned} Q(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)}) &= E\{\log f(y, \mathbf{u}; \boldsymbol{\Psi}) \mid \boldsymbol{\Psi}^{(s)}; \mathbf{y}\} \\ &= \sum_{i=1}^{n} \int \cdots \int \left[\log\left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i)\right] + \log g(\mathbf{u}_i; \boldsymbol{\Sigma})\right] h(\mathbf{u}_i \mid \boldsymbol{\Psi}^{(s)}; \mathbf{y}_i)\, d\mathbf{u}_i \end{aligned} \tag{7}$$

where $\mathbf{y}_i' = (\mathbf{y}_{i1}', \ldots, \mathbf{y}_{iT_i}')$. One can use Monte Carlo methods to approximate this expectation by generating samples from $h(\mathbf{u}_i \mid \boldsymbol{\Psi}^{(s)}; \mathbf{y}_i)$. We used the following multivariate rejection procedure (Geweke, 1996) to select K *iid* samples $\mathbf{z}_{il}^{(s)}$, $l = 1, \ldots, K$

(a) Sample $\mathbf{z}_{il}^{(s)}$ from $g(\mathbf{u}_i; \boldsymbol{\Sigma}^{(s)})$ and independently sample w from the uniform (0,1) distribution.

(b) Accept if $w \le \prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}^{(s)}; \mathbf{z}_{il}^{(s)})/\tau$, where $\tau = \sup_{\mathbf{u}} \prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}^{(s)}; \mathbf{u}_i)$; otherwise go to (a).

To calculate τ, one can regard $\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}^{(s)}; \mathbf{u}_i)$ as the likelihood for a multivariate GLM with unknown parameter vector $\mathbf{u}_i$. Thus, one can find the ML estimate of $\mathbf{u}_i$, say $\hat{\mathbf{u}}_i$, by fitting such a model that includes an offset of $\mathbf{Z}_{ij}\boldsymbol{\beta}^{(s)}$, and then calculate $\tau = \prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}^{(s)}; \hat{\mathbf{u}}_i)$. Given n sets of K multivariate samples from $h(\mathbf{u}_i \mid \boldsymbol{\Psi}^{(s)}; \mathbf{y}_i)$, $i = 1, \ldots, n$, the approximation to $Q(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)})$ is

$$Q_K(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)}) = \frac{1}{K}\sum_{i=1}^{n}\sum_{l=1}^{K}\left[\log\left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{z}_{il}^{(s)})\right] + \log g(\mathbf{z}_{il}^{(s)}; \boldsymbol{\Sigma})\right] \tag{8}$$

The M (maximization)-step at iteration $(s+1)$ consists of maximizing the Monte Carlo approximation $Q_K(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)})$ in (8) with respect to $\boldsymbol{\Psi}$. Since the elements $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ of $\boldsymbol{\Psi}$ occur separately in the two terms of (8), the terms can be maximized individually. The first term, $(1/K)\sum_i \sum_l \log[\prod_j f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{z}_{il}^{(s)})]$, refers to a likelihood for a multivariate GLM. By replicating the data vectors $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij}$ K times, the linear predictor for the lth multivariate sample of the jth observation on the ith subject can be expressed as

$$\boldsymbol{\eta}_{ijl} = \mathbf{Z}_{ijl}\boldsymbol{\beta} + \mathbf{W}_{ijl}\mathbf{z}_{il}^{(s)}$$

One can maximize with respect to $\boldsymbol{\beta}$ using Fisher scoring with $\mathbf{W}_{ijl}\mathbf{z}_{il}^{(s)}$ as an offset term. The second term to be maximized in (8) is just the log of a multivariate normal density. For unstructured covariance matrices or when the random effects are assumed independent, closed-form solutions exist for the ML estimate of $\boldsymbol{\Sigma}$.

The approximation of $Q(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)})$ in (8) inevitably contains Monte Carlo error. This error propagates through to the M-step, resulting in the estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ only approximating ML. Reducing the Monte Carlo error by increasing K also increases, of course, the time needed for the routine to converge. Booth and Hobert (1999) proposed a method for evaluating the Monte Carlo error in the estimates and automating the choice of K at each iteration to accurately evaluate the convergence. We adapted their method to this problem. Upon convergence we used the method of Louis (1982) to obtain estimated standard errors.

3.3 Pseudo-likelihood algorithm

Both the adaptive quadrature and automated Monte Carlo EM algorithms utilize numerical integration. Though based on approximations, in principle their estimates can be considered ML, since one can increase the Monte Carlo sample size or the number of quadrature points until the desired accuracy is obtained. In contrast, alternative fitting methods provide *approximate* ML estimates. Advantages are that they avoid the intractable integrals completely, thus being computationally much simpler, and they mimic methods for Gaussian linear mixed models. Articles that have proposed approximate ML estimation in GLMMs (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) have used Taylor series or Laplace approximations to avoid the intractable integrals. Keen and Engel (1997) proposed an iteratively re-weighted REML estimation routine for mixed ordinal regression models. Their method relates to the penalized quasi-likelihood aproach of Breslow and Clayton (1993). This subsection extends the PL approach of Wolfinger and O'Connell (1993) to provide approximate ML estimates for multinomial random effects models. The PL procedure iteratively fits a weighted Gaussian linear mixed model to a modified response vector.

Related to this, we note that there is not uniformity of opinion that forming (5) is an appropriate way to proceed with GLMMs. For instance, in generalizing the Henderson approach for normal mixed models to GLMMs with non-normal response and conjugate random effects distributions, Lee and Nelder (1996) defined an 'h-likelihood' that is a function of the random and fixed effects. In this and follow-up papers, they argued that advantages of their approach include avoiding the integration issue, treating the random effects as fixed but unknown quantities to be estimated like the fixed effects, producing a generalization of REML for GLMMs, and ready-made checking of the random-effect distribution. The discussion of Lee and Nelder (1996) provides an interesting debate on these issues.

To motivate the PL algorithm, we consider the model

$$\bar{\mathbf{y}} = \boldsymbol{\pi} + \mathbf{e} \tag{9}$$

for the complete response vector $\bar{\mathbf{y}} = [\bar{\mathbf{y}}_{ij}]$, with link function

$$\mathbf{g}(\boldsymbol{\pi}) = \mathbf{Z}\boldsymbol{\beta} + \mathbf{W}\mathbf{u}$$

where $\mathbf{u}' = (\mathbf{u}_1', \ldots, \mathbf{u}_n')$ with $\mathbf{u}_i$ being $\mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, and $\mathbf{e}' = (\mathbf{e}_{11}', \ldots, \mathbf{e}_{ij}', \ldots, \mathbf{e}_{nT_n}')$ have $E(\mathbf{e}_{ij} \mid \boldsymbol{\pi}_{ij}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{e}_{\mathrm{ij}} \mid \boldsymbol{\pi}_{\mathrm{ij}}) = \mathbf{R}_{\boldsymbol{\pi}_{\mathrm{ij}}}^{1/2}\, \mathbf{R}\, \mathbf{R}_{\boldsymbol{\pi}_{\mathrm{ij}}}^{1/2}$. For the multinomial models, $\mathbf{R}_{\boldsymbol{\pi}_{ij}} = [\mathrm{diag}(\boldsymbol{\pi}_{ij}) - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{ij}']/n_{ij}$, and following Wolfinger and O'Connell (1993) the covariance matrix $\mathbf{R}$ allows modelling of population-averaged associations. For the complete data model, we define $\mathrm{cov}(\mathbf{u}) = \boldsymbol{\Sigma}^c$ and $\mathrm{cov}(\mathbf{e} \mid \boldsymbol{\pi}) = \mathbf{R}_{\boldsymbol{\pi}}^{1/2}\, \mathbf{R}^c\, \mathbf{R}_{\boldsymbol{\pi}}^{1/2}$. First, for known estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$, let

$$\hat{\boldsymbol{\pi}} = \mathbf{h}(\hat{\boldsymbol{\eta}}) = \mathbf{h}(\mathbf{Z}\hat{\boldsymbol{\beta}} + \mathbf{W}\hat{\mathbf{u}})$$

where $\mathbf{h}(\cdot)$ is the inverse link for the desired model. Then, a Taylor series approximation to the residuals $\mathbf{e} = \bar{\mathbf{y}} - \boldsymbol{\pi}$ from (9) about the current estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ is given by

$$\begin{aligned}\mathbf{e} \approx \tilde{\mathbf{e}} &= (\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}}) - (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})' \frac{\mathrm{d}(\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}})}{\mathrm{d}\boldsymbol{\eta}} \\ &= (\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}}) - (\mathbf{Z}\boldsymbol{\beta} - \mathbf{Z}\hat{\boldsymbol{\beta}} + \mathbf{W}\mathbf{u} - \mathbf{W}\hat{\mathbf{u}})'\mathbf{D}\end{aligned} \tag{10}$$

where $\mathbf{D}$ is block diagonal with elements $\mathbf{D}_{ij} = \mathrm{d}\mathbf{h}/\mathrm{d}\boldsymbol{\eta}$ that depend on the model. Next, we approximate the conditional distribution of $(\tilde{\mathbf{e}} \mid \boldsymbol{\beta}, \mathbf{u})$ in (10) with a Gaussian distribution that has the same first two moments as $(\mathbf{e} \mid \boldsymbol{\beta}, \mathbf{u})$. Thus, we assume that

$$(\tilde{\mathbf{e}} \mid \boldsymbol{\beta}, \mathbf{u}) \sim N(\mathbf{0}, \mathbf{R}_{\boldsymbol{\pi}}^{1/2}\, \mathbf{R}^c\, \mathbf{R}_{\boldsymbol{\pi}}^{1/2}) \tag{11}$$

Then, using (10), (11), and approximating $\boldsymbol{\pi}$ in $\mathbf{R}_{\boldsymbol{\pi}}^{1/2}\, \mathbf{R}^c\, \mathbf{R}_{\boldsymbol{\pi}}^{1/2}$ with $\hat{\boldsymbol{\pi}}$, we obtain

$$\mathbf{D}^{-1'}(\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}}) \sim MVN[\mathbf{Z}\boldsymbol{\beta} - \mathbf{Z}\hat{\boldsymbol{\beta}} + \mathbf{W}\mathbf{u} - \mathbf{W}\hat{\mathbf{u}},\ \mathbf{D}^{-1}\mathbf{R}_{\hat{\boldsymbol{\pi}}}^{1/2}\mathbf{R}^c\mathbf{R}_{\hat{\boldsymbol{\pi}}}^{1/2}\, \mathbf{D}^{-1'}] \tag{12}$$

It follows from (12) that the approximate 'pseudo' observation vector,

$$\check{\mathbf{y}} = \mathbf{g}(\hat{\boldsymbol{\pi}}) + \mathbf{D}^{-1'}(\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}}) \tag{13}$$

has the approximate conditional distribution

$$(\check{\mathbf{y}} \mid \boldsymbol{\beta}, \mathbf{u}) \sim N[\mathbf{Z}\boldsymbol{\beta} + \mathbf{W}\mathbf{u},\ \mathbf{D}^{-1}\, \mathbf{R}_{\hat{\boldsymbol{\pi}}}^{1/2}\, \mathbf{R}^c\, \mathbf{R}_{\hat{\boldsymbol{\pi}}}^{1/2}\, \mathbf{D}^{-1'}] \tag{14}$$

Now, (14) has the form of a weighted linear mixed model with response $\check{\mathbf{y}}$ and weight matrix $\hat{\mathcal{W}} = \mathbf{D}'\, \mathbf{R}_{\hat{\boldsymbol{\pi}}}^{-1}\, \mathbf{D}$.

The log-likelihood corresponding to (14) is easily obtained. Following Wolfinger and O'Connell (1993), we insert an additional dispersion parameter ϕ and re-express the covariance matrices $\boldsymbol{\Sigma}^c$ and $\mathbf{R}^c$ as $\boldsymbol{\Sigma}^{c*} = \phi^{-1}\boldsymbol{\Sigma}^c$ and $\mathbf{R}^{c*} = \phi^{-1}\mathbf{R}^c$. The dispersion parameter is analogous to that used in quasi-likelihoods and can be equated to 1.0 if not needed. The resulting log-likelihood is

$$l(\boldsymbol{\beta}, \phi, \boldsymbol{\Sigma}^{c*}, \mathbf{R}^{c*}) = -\frac{1}{2}\log\ \mid \phi\mathbf{V} \mid -\frac{1}{2}\phi^{-1}(\check{\mathbf{y}} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{V}^{-1}(\check{\mathbf{y}} - \mathbf{Z}\boldsymbol{\beta}) - \frac{n}{2}\log(2\pi) \tag{15}$$

where

$$\mathbf{V} = \mathcal{W}^{-1/2}\mathbf{R}^{c*}\mathcal{W}^{-1/2} + W\boldsymbol{\Sigma}^{c*}\mathbf{W}' \tag{16}$$

Closed-form solutions for $\boldsymbol{\beta}$ and ϕ that maximize (15) exist and are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\hat{\mathbf{V}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}\hat{\mathbf{V}}^{-1}\check{\mathbf{y}} \tag{17}$$

and

$$\hat{\phi} = \frac{1}{n-p}\hat{\mathbf{r}}'\hat{\mathbf{V}}^{-1}\hat{\mathbf{r}} \tag{18}$$

where $\hat{\mathbf{r}} = \check{\mathbf{y}} - \mathbf{Z}(\mathbf{Z}'\,\hat{\mathbf{V}}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\,\hat{\mathbf{V}}^{-1}\,\check{\mathbf{y}}$. To obtain the estimates of $\boldsymbol{\Sigma}^{c*}$ and $\mathbf{R}^{c*}$ in $\hat{\mathbf{V}}$, the restricted profile likelihood

$$\begin{aligned} l_R(\boldsymbol{\Sigma}^{c*}, \mathbf{R}^{c*}) = & -\tfrac{1}{2}\log\,|\,\mathbf{V}\,| - \frac{n-p}{2}\log\,(\mathbf{r}'\mathbf{V}^{-1}\mathbf{r}) - \tfrac{1}{2}\log\,|\,\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\,| \\ & - \frac{n-p}{2}\{1 + \log\,[2\pi/(n-p)]\} \end{aligned}$$

can be maximized, providing REML-like estimates for $\boldsymbol{\Sigma}^{c*}$ and $\mathbf{R}^{c*}$. An estimate of $\mathbf{u}$ can then be found from $\hat{\mathbf{u}} = \hat{\boldsymbol{\Sigma}}^{c*}\,\mathbf{W}'\,\hat{\mathbf{V}}^{-1}\hat{\mathbf{r}}$. The PL algorithm consists of iteratively estimating $\boldsymbol{\beta}$ and $\mathbf{u}$, and $\boldsymbol{\Sigma}^c$ and $\mathbf{R}^c$, until convergence. Upon convergence, an estimate of the approximate covariance matrix for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ results from inverting

$$\begin{bmatrix} \mathbf{Z}'\hat{\mathcal{W}}^{1/2}\hat{\mathbf{R}}^{c^{-1}}\mathcal{W}^{1/2}\mathbf{Z} & \mathbf{Z}'\hat{\mathcal{W}}^{1/2}\hat{\mathbf{R}}^{c^{-1}}\mathcal{W}^{1/2}\mathbf{W} \\ \mathbf{W}'\hat{\mathcal{W}}^{1/2}\hat{\mathbf{R}}^{c^{-1}}\hat{\mathcal{W}}^{1/2}\mathbf{Z} & \mathbf{W}'\mathcal{W}^{1/2}\hat{\mathbf{R}}^{c^{-1}}\hat{\mathcal{W}}^{1/2}\mathbf{W} + \hat{\boldsymbol{\Sigma}}^{c^{-1}} \end{bmatrix}$$

PL algorithms are fast, since they avoid numerical integration. With GLMMs, however, these methods have been shown to be biased for highly non-normal cases, such as Bernoulli response data (Breslow and Clayton, 1993; Engel, 1998), with the bias increasing as variance components increase. We suspect that similar problems exist for the multinomial random effects models, both for estimating regression coefficients and variance components, when the multinomial sample sizes are small. Further research is needed to investigate the accuracy of the approximate methods for multinomial models. At the very least, their speed suggests these approximate methods can provide starting values for the ML algorithms considered previously and can be used in large-scale exploratory analyses.

3.4 Inference and prediction

Besides estimating parameters, one might also conduct inference about the parameters or obtain predictions for the random effects. One can conduct tests of the fixed effect parameters using the usual ML inferential techniques such as likelihood-ratio tests. Testing of variance components for the random effects is not straightforward. The likelihood-ratio statistic for testing that a single variance component is zero has null distribution that is approximately an equal mixture of degenerate at 0 and chi-squared on 1 df (Self and Liang, 1987). Score tests have ordinary chi-squared limiting distributions (Lin, 1997), but their performance for small to moderate samples with multivariate models is uncertain.

Prediction of the random effects $\mathbf{u}_i$, $i = 1, \ldots, n$, or linear combinations of fixed and random effects is based on the conditional expectation of $\mathbf{u}_i$ given the data and the final parameter estimates. For multinomial random effects models, this expectation has the form

$$E[\mathbf{u}_i \mid \mathbf{y}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}] = \frac{\int \cdots \int \mathbf{u}_i [\prod_{j=1}^{T_i} f(\mathbf{y}_{ij} \mid \hat{\boldsymbol{\beta}}; \mathbf{u}_i)]\, g(\mathbf{u}_i; \hat{\Sigma})\, \mathrm{d}\mathbf{u}_i}{\int \cdots \int [\prod_{j=1}^{T_i} f(\mathbf{y}_{ij} \mid \hat{\boldsymbol{\beta}}; \mathbf{u}_i)]\, g(\mathbf{u}_i; \hat{\Sigma})\, \mathrm{d}\mathbf{u}_i}$$

and requires integral approximations, for instance using adaptive Gauss–Hermite quadrature or Monte Carlo integration. Although the expectation involves only the data for cluster i, the estimates of $\{\mathbf{u}_i\}$ borrow information from all the clusters since $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ are obtained using all the data.

4 Model fitting: a semi-parametric approach

Advantages of the multivariate normality assumption for random effects include a variety of covariance structures and connections with Gaussian linear mixed models. An obvious concern of making this assumption is the unknown effect of misspecification of the random effects distribution. Some work exists on investigating these effects. Examining models for censored longitudinal economic data, Heckman and Singer (1984) showed that estimates of fixed parameters in a particular Weibull regression model were highly sensitive to misspecification. Less dramatic evidence occurs for other types of models (Davies, 1987; Neuhaus *et al*., 1992; Butler and Louis, 1992). In light of this evidence, some recent work has focused on non-parametric approaches (Heckman and Singer, 1984; Davies, 1987; Wood and Hinde, 1987; Follmann and Lambert, 1989; Butler and Louis, 1992; Wedel and DeSarbo, 1995; Aitkin, 1996; Aitkin, 1999). A referee has pointed out to us that this work has connections with a semi-parametric approach to estimation in the psychometric literature on latent trait and latent class models. See Heinen (1996, Chap. 6) for discussion.

This section considers a semi-parametric approach we have used with the ordinal random effects models. We outline an EM algorithm for obtaining non-parametric ML (NPML) estimates that generalizes Aitkin (1999), where the ‘non-parametric’ label here refers only to the lack of assumption about the random effects distribution. We refer to the full modelling approach as semi-parametric rather than non-parametric, since basic models have the same multinomial form as in the fully parametric case and since the non-parametric part of the analysis involves estimating its own parameters for the random effects distribution (namely, mass points and their probabilities). Like the approximate ML solutions in the previous section, it avoids the need for numerical integration.

4.1 NPML EM algorithm

For ease of notation, we develop the NPML EM algorithm for models with a single random effect, such as simple random intercept models. We assume that $g(u)$ is a discrete probability mass function with unknown finite support size K, mass points $\mathbf{m}' = (m_1, \ldots, m_K)$, and probabilities $\mathbf{p}' = (p_1, \ldots, p_K)$. For fixed K, the complete log-likelihood is

$$\log L(\mathbf{y}, \mathbf{u}; \boldsymbol{\beta}) = \sum_{i=1}^{n} \log \left[\prod_{j=1}^{T_i} f(\mathbf{y}_{ij} \mid u_i; \boldsymbol{\beta}) \right] + \sum_{i=1}^{n} \log g(u_i) \qquad (19)$$

where $\mathbf{y}$ and $\mathbf{u}$ denote the complete observed and unobserved data.

In the E-step at iteration $(s+1)$, the expectation of the complete log-likelihood (19) is calculated with respect to $f(\mathbf{u} \mid \mathbf{y}, \boldsymbol{\beta}^{(s)}, \mathbf{m}^{(s)}, \mathbf{p}^{(s)})$, where $(\boldsymbol{\beta}^{(s)}, \mathbf{m}^{(s)}, \mathbf{p}^{(s)})$ are estimates from the previous iteration. Using independence and Bayes Rule

$$\begin{aligned} &E\left[\log L(\boldsymbol{\beta}, \mathbf{m}, \mathbf{p} \mid \boldsymbol{\beta}^{(s)}, \mathbf{m}^{(s)}, \mathbf{p}^{(s)})\right] \\ &\quad = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{[\log f(\tilde{\mathbf{y}}_i \mid m_k, \boldsymbol{\beta}) + \log p_k]\, p_k^{(s)} f(\tilde{\mathbf{y}}_i \mid m_k^{(s)}, \boldsymbol{\beta}^{(s)})}{\sum_{l=1}^{K} p_l^{(s)} f(\tilde{\mathbf{y}}_i \mid m_l^{(s)}, \boldsymbol{\beta}^{(s)})} \\ &\quad = \sum_{i=1}^{n} \sum_{k=1}^{K} [\alpha_{ik}^{(s)} \log f(\tilde{\mathbf{y}}_i \mid m_k, \boldsymbol{\beta}) + \alpha_{ik}^{(s)} \log p_k] \end{aligned} \tag{20}$$

where

$$\alpha_{ik}^{(s)} = \frac{p_k^{(s)} f(\tilde{\mathbf{y}}_i \mid m_k^{(s)}, \boldsymbol{\beta}^{(s)})}{\sum_{l=1}^{K} p_l^{(s)} f(\tilde{\mathbf{y}}_i \mid m_l^{(s)}, \boldsymbol{\beta}^{(s)})}$$

represents the estimated posterior probability that the response vector $(\mathbf{y}'_{i1}, \ldots, \mathbf{y}'_{iT_i})$ for cluster i comes from component k.

The M-step consists of maximizing (20) with respect to $\boldsymbol{\beta}$, $\mathbf{m}$, and $\mathbf{p}$. The second term on the right of (20) is not a function of $\boldsymbol{\beta}$ or $\mathbf{m}$ and can be maximized separately from the first term, yielding $\hat{p}_k^{(s)} = \sum_{i=1}^{n} \alpha_{ik}^{(s)}/n$. Since $\{\alpha_{ik}^{(s)}\}$ are known, the first term on the right of (20) is the log-likelihood of a weighted multivariate GLM with known weights. Wood and Hinde (1987) noted that $\{m_k\}$ can be estimated by incorporating a K-level factor in the model in place of m_k. By adjusting the model and response matrices accordingly, one can maximize the weighted multivariate GLM using a Fisher scoring algorithm. The E-step and the M-step are then iterated until convergence in the parameter estimates and the deviance. To determine K, we recommend successively fitting the NPML EM algorithm while increasing K until no further improvement in the deviance is obtained. Usually K is quite small and the estimate of the true mixture distribution may be quite poor. Some find this unappealing and prefer a smooth mixing density (Davidian and Gallant, 1993). Based on simulation results summarized later in the paper, it appears that the poor estimation of the mixture distribution does not cause problems with estimating the fixed effects. Thus, based on our experience it seems to us that this approach is usually reasonable when main interest focuses on the fixed effects rather than the mixture distribution itself.

Upon convergence, the inverse of the observed information matrix provides standard errors for the parameter estimates (Hartzel, 1999). For starting values for the mass points and their probabilities, we used the nodes and weights from Gauss–Hermite quadrature. However, it is advisable to try a few sets of starting values. As noted by Wood and Hinde (1987) and from our own experience, the estimated mixing distribution can place positive probability at $\pm\infty$. We have observed this with ordinal models when observations in many clusters all have response in the highest category or all in the lowest category. Thus, routinely we also fit the model with mass points at $\pm\infty$ (as in Wood and Hinde, 1987) and evaluate the log-likelihood. (For an ordinal model with random intercept, a mass point at $+\infty$ $(-\infty)$ corresponds to a probability of 1 that the response falls at the lowest (highest) level. The fitting of models with such mass points therefore amounts to putting conditions in the code that ignores covariates and sets the multinomial likelihood to one or zero when the random effect is at ±∞.) When the ML fit has such mass points, in our experience this does not affect estimates of $\boldsymbol{\beta}$ or their standard errors in the ordinal models (2) and (3), but not surprisingly it does affect estimates of intercept parameters.

4.2 Inference and prediction

Ideally, the asymptotic inferential theory for the estimation of fixed effects would parallel the standard ML theory, but such theory is lacking. The difficulty arises partly from the unknown mixture support size K for a model or for each of two models to be compared. This implies the dimension of the parameter vector is unknown. Despite this, recommendations have relied on standard ML theory. For instance, the likelihood-ratio test has been used for testing fixed parameters and making model comparisons (Davies, 1987; Wood and Hinde, 1987; Aitkin, 1996). Hartzel (1999) examined the Wald and likelihood-ratio tests for multinomial logit random effects models and concluded that they provided reasonably appropriate inference for the NPML approach.

Work is also needed on inferences about the mixture distribution g. For instance, one might want to compare a model to one not containing the random effect. This entails testing that the masses are on the boundary of the parameter space, which precludes the use of the likelihood-ratio test. The same comment applies to comparing NPML fits with different numbers of mass points.

A final issue is model identifiability. A given mixture is identifiable if it is uniquely characterized in the sense that two distinct sets of parameters defining the mixture cannot yield the same distribution. Teicher (1963) and Teicher (1967) provided conditions under which mixtures of binomial distributions are identifiable. Follmann and Lambert (1991) extended this to mixtures of logistic regression models with random intercepts. Butler and Louis (1997) considered a latent linear model for binary data with any class of mixing distribution and provided sufficient conditions for identifiability of the fixed effects and mixing distribution. As in the binary case, not all mixtures of multinomial models are identifiable. Hartzel (1999) provided sufficient conditions for the identifiability of overdispersed multinomial regression models; however, further work is needed for the more general multinomial random effects model considered here.

4.3 Effect of choice of random effect distribution

Having mentioned the advantage the semi-parametric approach has of avoiding potential misspecification of the random effects distribution brings up the issue of potential bias in parameter estimation with a misspecified parametric approach. Some research has studied the robustness of the parametric approach under different random effects distributions (Neuhaus *et al.*, 1992; Butler and Louis, 1992). These studies concentrated on binary logistic regression and indicated that the parametric ML approach was robust to misspecification of the mixing distribution in terms of estimating fixed effects but less so in estimating the mixing distribution and its characteristics.

We conducted a limited simulation study to compare NPML to parametric ML approaches in terms of the bias between the two approaches for a single covariate, random intercept ordinal logit model using a variety of different random effects distributions. We generated clustered response data according to a cumulative logit model

$$\operatorname{logit}(\pi_{ij1} + \cdots + \pi_{ijr}) = \alpha_r + \beta x_{ij} + u_i \quad r = 1, \ldots, R-1 \tag{21}$$

where $R = 3, i = 1, \ldots, 100$, and $j = 1, \cdots, T$ for $T = 4$ or 7. For all simulations, $\{x_{ij}\}$ were generated from the standard normal distribution, and $(\alpha_1, \alpha_2, \beta) = (-1, 1, 0.5)$. We let u_i be $N(0, 0.5)$, Exp(1), $U(-0.5, 0.5)$, discrete with mass points $(-0.5, 0.5)$ and masses (0.5, 0.5), or degenerate. For each case except the last, the random effect was standardized to have mean 0 and variance 0.5. For each combination, we simulated 500 datasets (this number reflecting computational constraints) and fit them with both the NPML algorithm and the adaptive quadrature algorithm.

As in Neuhaus *et al.* (1992), we observed only small estimated bias for the parametric ML estimation of β, even when the mixture distribution was very skewed. The largest estimated biases could be explained by Monte Carlo error. They corresponded to a percent bias of approximately 2.5%, which is similar to values Neuhaus *et al.* (1992) reported. For estimation of the intercept parameters, the largest estimated biases were only on the order of 2%. The parametric approach provided accurate estimates of the variance component σ^2 with a normal random effect. Considerable bias occurred for the remaining distributions, however. In general, larger estimated biases were found with a smaller cluster size.

For the semi-parametric approach, the estimated bias for β was similar to the parametric approach. This held even when simulated datasets had estimated mass points at $\pm\infty$. Estimation of the intercept parameters and the variance component was much less accurate, because these parameters are related to the mass points. As others have cautioned, one should not trust estimates that are functions of the mixing distribution estimates.

For each distribution for the random effects, standard errors of the fixed effect estimate were very similar for the parametric and semi-parametric approaches. In addition, for both approaches there was good agreement between the mean of the model-based estimated standard errors (using the observed information matrix for each model fit) and the estimated standard error based on the set of Monte Carlo simulations.

Conclusions from these simulations are tentative, because of their limited scope. They suggest that parametric and semi-parametric approaches produce essentially unbiased estimates of β, with similar behaviour under the various random effects distributions and cluster sizes. For estimation of the remaining parameters, the parametric approach seems more reliable, although it also had some difficulties when the random effects distribution was extremely skewed. We are currently conducting a more thorough simulation study of the effect of misspecification of the random effects distribution in generalized linear mixed models. A referee has suggested also including a covariate that varies only with i, and this is one design we plan to use in this study.

5 Movie critics example

Each week *Variety* magazine, in an article titled *Crix' Picks*, summarizes reviews of new movies by several critics. Each review is categorized as Pro (positive), Con (negative), or Mixed (a mixture of the two). Table 1 summarizes reviews of 93 movies from April, 1995 through March, 1997 by four critics. We compare the reviewers' ratings using multinomial logit models with random intercepts. Let π_{ijr} denote the probability that the rating of movie

Table 1 Ratings of movies by four movie critics

Lyons	Siskel	Medved = Pro			Medved = Mixed			Medved = Con		
						Ebert rating				
		Pro	Mixed	Con	Pro	Mixed	Con	Pro	Mixed	Con
Pro	Pro	15	2	0	2	0	0	8	0	2
	Mixed	0	3	0	2	3	0	3	2	1
	Con	1	0	1	1	2	1	1	1	2
Mixed	Pro	2	0	0	2	1	0	0	1	0
	Mixed	1	1	0	0	1	0	0	1	0
	Con	1	0	0	0	0	0	0	1	1
Con	Pro	3	1	1	0	0	0	4	1	0
	Mixed	0	0	1	0	0	0	1	3	0
	Con	0	0	1	2	1	0	2	2	4

Source: Data taken from *Variety*, April, 1995 through March, 1997 (thanks to Larry Winner for providing this).

i by critic j ($1 = \text{Siskel}$, $2 = \text{Ebert}$, $3 = \text{Lyons}$, $4 = \text{Medved}$) falls in category r. We obtained ML estimates using adaptive Gauss–Hermite quadrature. Reported estimates are based on 50 quadrature points, but most estimates and standard errors stabilized to the third decimal place after 10 points.

5.1 Parametric random effects

An ACL model is

$$\log\left(\frac{\pi_{ijr}}{\pi_{ij,r+1}}\right) = \alpha_r + \beta_j + u_{ir} \quad r = 1, 2 \tag{22}$$

where $\{u_{ir}\}$ are *iid*. For identifiability, we take $\beta_4 = 0$. Table 2 (in column 3 of the 7 models) shows $\{\hat{\beta}_j\}$ assuming multivariate normality for (u_{i1}, u_{i2}) with unspecified correlation and possibly different variances. Compared with Medved, all effects are positive, reflecting greater tendency for the other critic to have rating in the more positive category. For instance, for a given movie the estimated odds that Siskel's rating was pro instead of mixed, or mixed instead of con, are $\exp(0.52) = 1.7$ times the corresponding odds for Medved. Table 2 also shows in columns 4 and 6 estimates for the simpler random effects structure $u_{ir} = u_i$

(i) assuming a normal distribution for $\{u_i\}$
(ii) assuming $\sigma = 0$, in which case the four responses are mutually independent.

The $\{\hat{\beta}_j\}$ are similar except when $\sigma = 0$; they are then smaller, as they correspond to population-averaged estimates.

Table 2 also shows likelihood-ratio goodness-of-fit statistics comparing counts to their fitted values. The G^2 statistics are only rough indicators because of the sparseness. A cell-wise residual analysis indicates model (22) fits reasonably well, but a model discussed below indicates there is some lack of fit. In the parametric case, the more complex random effects structure provides a better fit, but does not provide substantively different results about the fixed effects.

Table 2 ML estimates for ACL random effects model (24) with different logit effects β^*_{jr} and model (22) with common effects β_j for movie critics

Effect	Model						
	(24)*	(24)* Corr = 0	(22)	(22) $u_{ir} = u_i$	(22) $u_{ir} = u_i$ nonpar	(22) $u_{ir} = 0$ ($\sigma = 0$)	(25) ord. quasi symmetry
β_1 (Siskel)	0.095, 0.965	0.144, 0.958	0.519	0.520	0.526	0.381	0.522
se	0.439, 0.460	0.427, 0.450	0.201	0.201	0.203	0.170	0.203
β_2 (Ebert)	0.081, 1.806	0.163, 1.790	0.854	0.854	0.860	0.630	0.855
se	0.433, 0.499	0.418, 0.489	0.213	0.212	0.214	0.176	0.214
β_3 (Lyons)	1.002, 0.194	1.002, 0.246	0.640	0.641	0.647	0.471	0.643
se	0.477, 0.493	0.466, 0.481	0.205	0.205	0.206	0.172	0.206
SD(u_{i1})	1.41	1.24	1.40	0.80	∞		
SD(u_{i2})	1.45	1.27	1.31	–	–		
Corr(u_{i1}, u_{i2})	−0.39	0.0	−0.34	–	–		
G^2	72.0	72.7	80.6	90.8	86.6	118.9	73.4
d*f*	69	70	72	74	69	75	63

*First estimate and standard error refers to effect β^*_{j1} for $\log(\pi_{ij1}/\pi_{ij2})$, and second to effect β^*_{j2} for $\log(\pi_{ij2}/\pi_{ij3})$.

A baseline-category logit model for Table 1 is

$$\log\left(\frac{\pi_{ijr}}{\pi_{ij3}}\right) = \alpha_r + \beta_{jr} + u_{ir} \quad r = 1, 2 \tag{23}$$

with $\beta_{4r} = 0$ for each r and unstructured covariance matrix for (u_{i1}, u_{i2}). This corresponds to an ACL model with differing rater effects for each logit

$$\log\left(\frac{\pi_{ijr}}{\pi_{ij,r+1}}\right) = \alpha^*_r + \beta^*_{jr} + u_{ir} \quad r = 1, 2 \tag{24}$$

where $\beta^*_{j1} = \beta_{j1} - \beta_{j+1,2}$ and $\beta^*_{j2} = \beta_{j2}$ and again the covariance matrix is unstructured. Fitting the model in this form makes it easier to detect ways in which model (22) with common effect for each logit may have lack of fit. Table 2 (in column 1) showing ML estimates for this model provides evidence of heterogeneity in the effects. The estimates show a negligible effect comparing Siskel and Ebert to Medved for the pro versus mixed comparison and a negligible effect comparing Lyons to Medved for the mixed versus con comparison. The model has a reduction in G^2 of 8.6 (df = 3) compared to the simpler ACL model. None of the cell-specific residuals indicated problems with lack of fit. Results are substantively similar for the model with zero correlation between the random effects (in column 2).

5.2 Semi-parametric random effects approaches

Column 5 of Table 2 shows results using a non-parametric treatment of $\{u_i\}$. Estimates $\{\hat{\beta}_j\}$ are very similar to those in Column 4 for the corresponding model with normal random effects. For the semi-parametric approach, Table 3 shows how the estimates and standard errors change as the number of mass points increases. Since infinite mass points occur for

Table 3 Estimates of critic effects for ACL model with non-parametric random effects applied to Table 1

	Number of mass points				
	1	2	3	4	5
Log-likelihood	−379.5	−366.6	−363.7	−363.4	−363.4
β_1 (Siskel)	0.381	0.508	0.522	0.526	0.526
se	0.170	0.199	0.202	0.203	0.203
β_2 (Ebert)	0.630	0.828	0.854	0.860	0.860
se	0.176	0.208	0.212	0.214	0.214
β_3 (Lyons)	0.471	0.625	0.654	0.647	0.647
se	0.172	0.201	0.205	0.206	0.206
G^2	118.9	93.0	87.3	86.6	86.6
df	75	73	71	69	69
Mass 1, prob	0.427, 1	−0.474, 0.487	−1.070, 0.201	$-\infty$, 0.024	$-\infty$, 0.024
Mass 2, prob		0.833, 0.512	−0.150, 0.677	−0.704, 0.277	−0.704, 0.277
Mass 3, prob			$+\infty$, 0.122	0.223, 0.581	0.223, 0.395
Mass 4, prob				$+\infty$, 0.118	0.223, 0.186
Mass 5, prob					$+\infty$, 0.118

some K, we set $\hat{\alpha}_1 = 0$ and impose no restriction on the mass points. With $K = 1$, the model is equivalent to the parametric model with $\sigma = 0$. The model with $K = 2$ provides a substantial improvement in fit. With $K = 3$, an infinite mass point occurs, having probability 0.12. The log likelihood is essentially constant (to the fifth decimal place) when the third mass point exceeds roughly 15. That mass point forces at least 12% of the joint distribution in the cell whereby each critic's rating is pro (16% of the sample is in that cell). With $K = 4$, there is also a small mass at $-\infty$, forcing at least 2% of the joint distribution in the cell whereby each critic's rating is con (4% of the sample is in that cell). The fit with $K = 4$ is the ML fit, as allowing $K = 5$ gives essentially the same fit and only four distinct mass points.

An interesting connection exists between a loglinear model and a semi-parametric approach with ACL random effects models for within-subjects effects. We illustrate with model (22) for the movie critic data. Let $\{n(r_1, r_2, r_3, r_4)\}$ denote the cell counts in Table 1, and let $\mu(r_1, r_2, r_3, r_4) = E[n(r_1, r_2, r_3, r_4)]$ denote the expected frequency for response r_j by critic j, $j = 1, \ldots, 4$. It follows from Agresti (1993) that regardless of the random effects distribution in the ACL model (22), $\{\mu(r_1, r_2, r_3, r_4)\}$ satisfy the loglinear model

$$\log \mu(r_1, r_2, r_3, r_4) = \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3 + \rho(r_1, r_2, r_3, r_4) \qquad (25)$$

where $\rho(r_1, r_2, r_3, r_4)$ is permutation invariant. When model (22) holds, fitting this loglinear model provides another semi-parametric way to estimate consistently $\{\beta_j\}$. Table 2 also shows results for it. The ML estimates for the random effects models and the loglinear model are similar, approximately $\hat{\beta}_1 = 0.52$, $\hat{\beta}_2 = 0.85$, $\hat{\beta}_3 = 0.64$. Again, an advantage is the lack of assumption about the parametric form for the random effects distribution. A disadvantage is the lack of information provided about their variability. In fact, the loglinear $\{\hat{\beta}_j\}$ are necessarily identical to the conditional ML estimates from treating $\{u_{ij}\}$ as fixed effects and conditioning on their sufficient statistics.

Model (25) is a special case of the loglinear *quasi-symmetry model*, having a symmetric interaction term but allowing heterogeneous main effects. The ordinary quasi-symmetry model treats the response as nominal, but model (25) treats it as ordinal by using main effects that are quantitative variates instead of qualitative factors. Similar connections exist between baseline-category logit random effects models and ordinary quasi-symmetry models, see Conaway (1989). This more general model has $G^2 = 63.3, \mathrm{df} = 60$, showing some improvement over model (25). For the parametric random effects modelling or loglinear modelling, allowing differing effects for each logit reduces the deviance by about 10 on $\mathrm{df} = 3$.

6 Final comments

The semi-parametric approach to fitting multinomial logit random effects models has several unresolved issues. Based on the simulation results mentioned in Section 4.3, one might consider always using the normal parametric approach. Indeed, we have not observed cases in which the semi-parametric approach provided substantively different results. Therefore, we view the NPML approach mainly as a type of check for normal random effects modelling; if one obtains different results, it suggests further investigation into the model adequacy.

Part of the work for the thesis on which this paper is based (Hartzel, 1999) developed programs for fitting multinomial logit random effects models with the three strategies discussed in Section 3 and the semi-parametric approach. Recently, the SAS procedure NLMIXED has become available for ML fitting of GLMMs. Use of SAS for fitting the parametric random effects models to Table 1 is shown on the journal's webpage for this article (http://stat.uibk.ac.at/SMIJ).

Acknowledgements

This work was partially supported by grants from NIH and NSF. The authors appreciate many helpful comments from two referees, in particular pointing out several references and software of which we were unaware.

References

Adams J, Wilson M, Wang W (1997) The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement* **21**, 1–23.

Adams RJ, Wilson MR (1996) Formulating the Rasch model as a mixed coefficients multinomial logit. In: Engelhard G, Wilson MR, eds. *Objective measurement: theory and practice*. Vol. 3, Norwood, NJ: Ablex, 143–66.

Agresti A (1993) Computing conditional maximum likelihood estimates for generalized Rasch models using simple loglinear models with diagonals parameters. *Scandinavian Journal of Statistics* **20**, 63–71.

Aitkin M (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**, 251–62.

Aitkin M (1999) A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55**, 117–28.

Anderson DA, Aitkin M (1985) Variance component models with binary response: interviewer variability. *Journal of Royal Statistical Society, Series B* **47**, 203–210.

Bock RD (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* **37**, 29–51.

Booth JG, Hobert JP (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B* **61**, 265–85.

Breslow N (1984) Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38–44.

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Bryk AS, Raudenbush SW, Congdon R (2000) *HLM version 5*. Chicago: Scientific Software International, Inc.

Butler SM, Louis TA (1992) Random effects models with nonparametric priors. *Statistics in Medicine* **11**, 1981–2000.

Butler SM, Louis TA (1997) Consistency of maximum likelihood estimators in general random effects models for binary data. *The Annals of Statistics* **25**, 351–77.

Chen Z, Kuo L (2001) A note on the estimation of the multinomial logit model with random effects. *American Statistician* **55**, 89–95.

Chintagunta PK, Jain DC, Vilcassim NJ (1991) Investigating heterogeneity in brand preferences in logit models for panel data. *Journal of Marketing Research* **28**, 417–28.

Conaway M (1989) Analysis of repeated categorical measurements with conditional likelihood methods. *Journal of the American Statistical Association* **84**, 53–62.

Conaway MR (1990) A random effects model for binary data. *Biometrics* **46**, 317–28.

Coull B, Agresti A (2000) Random effects modeling of multiple binary responses using the multivariate binomial logit-normal distribution. *Biometrics* **56**, 162–68.

Crouchley R (1995) A random-effects model for ordered categorical data. *Journal of the American Statistical Association* **90**, 489–98.

Daniels MJ, Gatsonis C (1997) Hierarchical polytomous regression models with applications to health services research. *Statistics in Medicine* **16**, 2311–25.

Davidian M, Gallant AR (1993) The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80**, 475–88.

Davies RB (1987) Mass point methods for dealing with nuisance parameters in longitudinal studies. In: Crouchley R, ed. *Longitudinal data analysis*. Avebury, Aldershot: Hants, 88–109.

Dempster AP, Laird NM, Rubin DA (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Dos Santos D, Berridge D (2000) A continuation ratio random effects model for repeated ordinal responses. *Statistics in Medicine* **19**, 3377–88.

Engel B (1998) A simple illustration of the failure of PQL, IRREML and APHL as approximate ML methods for mixed models for binary data. *Biometrical Journal* **40**, 141–54.

Ezzet F, Whitehead J (1991) A random effects model for ordinal responses from a crossover trial. *Statistics in Medicine* **10**, 901–07.

Fahrmeir L, Tutz G (2001) *Multivariate statistical modelling based on generalized linear models*. 2nd edn. New York: Springer-Verlag.

Follmann DA, Lambert D (1989) Generalizing logistic regression by nonparametric mixing. *Journal of the American Statistical Association* **84**, 295–300.

Follmann DA, Lambert D (1991) Identifiability of finite mixtures of logistic regression models. *Journal of Statistical Planning and Inference* **27**, 375–81.

Geweke J (1996) *Handbook of computational statistics*. New York: Elsevier, Chapter 15.

Gilmour AR, Anderson RD, Rae AL (1985) The analysis of data by a generalized linear mixed model. *Biometrika* **72**, 593–99.

Goldstein H, Rasbash J, Plewis I, Draper D, Browne W, Wang M (1998) *A user's guide to MLwiN*. University of London. London: Institute of Education.

Gönül F, Srinivasan K (1993) Modeling unobserved heterogeneity in multinomial logit models: methodological and managerial implications. *Marketing Science* **12**, 213–29.

Greene WH (1998) *LIMDEP version 7.0 user's manual*. Plainview, NY: Econometric Software, Inc.

Hartzel, J (1999) *Random effects models for nominal and ordinal data*. PhD thesis, University of Florida.

Hartzel J, Liu I, Agresti A (2001) Describing heterogeneous effects in stratified ordinal contingency tables, with application to multi-centre clinical trials. *Computational Statistics and Data Analysis*.

Harville DA, Mee RW (1984) A mixed-model procedure for analysing ordered categorical data. *Biometrics* **40**, 393–408.

Heckman JJ, Singer B (1984) A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrica* **52**, 271–320.

Hedeker D (1999) MIXNO: a computer program for mixed-effects nominal logistic regression. *Journal of Statistical Software* **4**, 1–92.

Hedeker D, Gibbons RD (1994) A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**, 933–44.

Hedeker D, Mermelstein RJ (1998) A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research* **33**, 427–55.

Heinen T (1996) *Latent class and discrete latent trait models: similarities and differences*. Sage.

Hinde JP (1982) Compound regression models. In: Gilchrist R, ed. *GLIM 82: international conference for generalized linear models*. New York: Springer, 109–121.

Jain DC, Vilcassim NJ, Chintagunta PK (1994) A random-coefficient logit brand-choice model applied to panel data. *Journal of Business and Economic Statistics* **12**, 317–28.

Jansen J (1990) On the statistical analysis of ordinal data when extravariation is present. *Applied Statistics* **39**, 74–85.

Keen A, Engel B (1997) Analysis of a mixed model for ordinal data by iterative reweighted REML. *Statistica Neerlandica* **51**, 129–44.

Lee Y, Nelder JA (1996) Hierarchical generalized linear models (disc: P656-678). *Journal of the Royal Statistical Society, Series B, Methodological* **58**, 619–56.

Lin X (1997) Variance component testing in generalised linear models with random effects. *Biometrika* **84**, 309–26.

Liu Q, Pierce DA (1994) A note on Gauss-hermite quadrature. *Biometrika* **81**, 624–29.

Louis TA (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226–33.

McCulloch CE (1994) Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association* **89**, 330–35.

McCulloch CE (1997) Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–70.

Neuhaus JM, Hauck WW, Kalbfleisch JD (1992) The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika* **79**, 755–62.

Pinheiro JC, Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

Rasch G (1961) On general laws and the meaning of measurement in psychology. In: Neyman J, ed. *Proceedings of the 4th Berkeley symposium on mathematical statistics and probability, vol 4*. Berkeley: University of California Press.

Revelt D, Train K (1998). Mixed logit with repeated choices: household choices of appliance efficiency level. *Review of Economics and Statistics* **80**, 647–57.

Self S, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–10.

Stiratelli R, Laird NM, Ware JH (1984) Random effects models for serial observations with binary responses. *Biometrics* **40**, 961–71.

Teicher H (1963) Indentifiability of finite mixtures. *Annals of Mathematical Statistics* **34**, 1265–69.

Teicher H (1967) Indentifiability of mixtures of product measures. *Annals of Mathematical Statistics* **38**, 1300–302.

Ten Have TR (1996) A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics* **52**, 473–91.

Ten Have TR, Uttal DH (1994) Subject-specific and population–averaged continuation ratio logit models for multiple discrete time survival profiles. *Applied Statistics* **32**, 371–84.

Tutz G, Hennevogl W (1996) Random effects in ordinal regression models. *Computational Statistics and Data Analysis* **22**, 537–57.

Wedel M, DeSarbo WS (1995) A mixture likelihood approach for generalized linear models. *Journal of Classification* **12**, 21–55.

Wolfinger R, O'Connell M (1993) Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233–43.

Wood A, Hinde J (1987) Binomial variance component models with a non-parametric assumption concerning random effects. In: Crouchley R, ed. *Longitudinal data analysis*. Avebury, Aldershot: Hants, 110–28.

Zeger SL, Karim MR (1991) Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79–86.